DOCUMENT RESUME

ED 412 746 FL 024 831

AUTHOR Elinor, Siaegh-Haddad
TITLE Reading Native and Foreign Language Texts and Tests: The Case of Arabic and Hebrew Native Speakers Reading L1 and English FL Texts and Tests.
PUB DATE 1997-00-00
NOTE 20p.; Paper presented at the Language Testing Symposium (ACROLT) (1997).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Arabic; College Students; Comparative Analysis; Contrastive Linguistics; English (Second Language); Foreign Countries; *Hebrew; Higher Education; Language Research; *Language Tests; *Multiple Choice Tests; Native Speakers; Scores; *Second Languages; *Test Items; Test Validity; Testing; Uncommonly Taught Languages
IDENTIFIERS Haifa University (Israel); Israel; *Open Ended Questions

ABSTRACT

A study investigated the effect of test item type (multiple-choice or open-ended) on performance reading comprehension tests given in both the student's native language and a second language. Subjects were 24 native Arabic-speaking and 38 native Hebrew-speaking students at Haifa University (Israel), all enrolled in a course in English as a second language. English language texts were selected from an Israeli standardized test, with two test item versions: multiple-choice and open-ended. Texts in Arabic and Hebrew were drawn from practice books designed to prepare students for a psychometric test, and similarly, two types of test item were prepared for each. All texts were controlled for readability, length, and neutrality of topic. For each test, reading processes were examined using the paraphrase/translation segment of a think-aloud protocol. Results across test context (test items vs. paraphrase/translation), across language type (English as a second language vs. native language), and native language group (Arabic vs. Hebrew) are examined, and implications concerning the construct validity of reading comprehension tests are discussed. Contains 19 references. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Author: Saiegh-Haddad Elinor

ED 412 746

•READING NATIVE AND FOREIGN LANGUAGE TEXTS AND TESTS- THE CASE OF ARABIC AND HEBREW NATIVE SPEAKERS READING L1 AND ENGLISH FL TEXTS AND TESTS.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Saiegh Haddad Elinor

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Introduction

Attempts at the unfolding of the reading comprehension construct have always been hindered by the inevitably indirect measurement of this trait. Because language ability, in general, is an unobservable, theoretical construct, it cannot be measured but indirectly through test questions. In other words, as Spolsky (1990) puts it, "Testers are concerned with performance, but aim to understand the underlying abilities and knowledge that can be revealed by performance (p.5). In a reading comprehension test, this indirect measurement and the amount of error it introduces into the reading test performance, is exacerbated by the fact that the performance of reading itself is also unobservable, and therefore requires of readers the *demonstration* of the ability to read by performing on a variety of testing tasks allegedly tapping that ability. This indirect measurement, coupled with the demonstration dimension inherent in the testing of the receptive skills, places a heavy burden on the testing tasks and methods and makes very difficult the distinction between the true, reading variance of reading comprehension test performance from the error, among others, measurement variance of it. This, as the main source of error in reading test performance highlighted the urgency to investigate the nature of the interaction between the various testing tasks and the specific abilities they measure for the sake of construct validation.

FL024831

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

The awareness of the role of task, or method, characteristics to test performance has , initially, resulted in the validation of tests, thorough a comparison of item content and

BEST COPY AVAILABLE

function with item performance. Such *a priori* validation, with its basic presupposition that the performance on the different items can be made predictable by means of item content has proved insufficient and misleading, due to two major reasons. Firstly, with the recognition of the role that the reader plays in the reading process, and the amount of information and experience he brings to the task of reading, empirical research proved that different readers perform differently on identical tasks. Moreover, the same reader was found to perform the same item differently, on different occasions (Messick 1989). Secondly, ample evidence was provided which showed that reading experts are incapable of, unanimously, predicting the level of difficulty of the different items, or the specific abilities they measure. Alderson and Lukmani (1989), and Alderson (1990), investigating the judgments of EFL reading experts about the kind of skills items were measuring, found that judges tended to disagree as to what each item was measuring, or to assign a specific skill to a specific item. In addition, they found no relationship between item statistics, namely, difficulty and discrimination, and the skill it was allegedly testing. Freedle and Kostin (1993) and Bachman et al (1995), examining the relationship between item statistics and item difficulty, reached a similar conclusion, namely, that variation in item difficulty cannot be appropriately accounted for by variation in item content characteristics.

Once it was found that the test content was not a sufficient source of information for test score interpretation, the construct validation of reading tests was no longer confined to a comparison between item content and item performance, but emphasized the inclusion of information about how testees tackle the testing task, and relating that to the test content and performance. Nevo (1989), for instance, investigated the kinds of strategies test-takers report while taking a multiple-Choice test, in the native and foreign language, and concluded that the stimulus format of the test, namely the text and questions, and the response format,

i.e., the alternatives, affect the strategies employed in taking the test, in both languages. Similarly, Storey (1994), guided by the same interest in describing the reading behavior that characterize the taking of different item techniques, examines the strategies employed in the taking of five different item types, frequently employed in the assessment of reading comprehension, and found that none of the item types investigated elicited, what he calls, 'authentic reading' (p.137), rather, each of them was found to generate test-specific reading behavior.

A further insight into the performance on Multiple-Choice (MC) items, and its implications to the construct validity of reading tests, was provided by Farr, Pritchard, and Smitten (1990), who used introspective and retrospective methods in gathering information on the MC test-taking strategies, and found that readers were driven, in their reading process, by a focus on finding the best response to each question, and did not demonstrate, what they called, 'typical reading'. The implications of this finding to the construct validity of MC tests were not regarded intimidating to the construct validity of the test, because, as they claim, these are the strategies used by readers who search a text for specific information, and because even though readers were not engaged in typical reading, they "had to understand what they were reading in order to understand the questions and to search for responses" (p.223).

The foregoing suggests that, though the issue of the reading test construct validation has always been an attractive domain of research, owing to the potential significance it has to the understanding of the reading process, the implications of such studies have varied widely, from the reiteration of the error, test-specific reading behavior triggered in the reading test performance, and which casts doubt on the construct validity of reading tests (Cohen 1984,

Anderson et al. 1985, Nevo 1989, and Storey 1994), to the acceptance of such test-specific reading behavior, that particular items do elicit, because, though atypical, and seemingly invalid, it does underlie reading comprehension (Farr, Pritchard, and Smitten 1990).

In this uncertainty about the relationship between item content, item performance, and the ramifications of that to the construct validity of Reading Comprehension tests, Anderson, Bachman, Perkins, and Cohen (1991) conducted a very valuable study on the necessity of considering multiple sources of information in construct validation. They combine information on test-taking strategies, item content, and item performance, in the construct validity of a reading comprehension test. This triangulation of sources of information rendered a significant relationship between the type of strategies used and the type of question being asked, which was also found to fluctuate when the content of the items is examined through a different paradigm. In addition, no significant correlation was found between the item type, as determined by the test designers, and the subsequent item difficulty. Such item-specific reading behavior, which was not consistently predictable by means of item type, has highlighted the value of the adopted triangulation approach to construct validation, which, as the authors rightly point out, is "perhaps the greatest insight gained from this investigation" (p.61).

Guided by the same interest in exploring the nature of the reading performance and process in the test-taking context, and accepting the recommendation made in Anderson and colleagues (1991), regarding the use of several data sources in the validation of reading tests, the currently reported research set out to investigate the reading test performance and process, by contrasting the reading of a text with the reading of a test, i.e., a text plus questions, when these questions are presented in two item types: Multiple-Choice and Open-

Ended. In addition, in order to examine potentially native-language-dependent reading behavior, and the possible transfer of reading behavior, from the mother tongue to the foreign language reading context, the non-test reading vs. the test reading was investigated with native speakers of Arabic and Hebrew, reading native and foreign language texts and tests. Information on the construct validation of the reading tests used was sought, initially, in the performance of subjects in the different reading contexts, text vs. test, and later, in the process followed by the individual readers. The reading process was measured by two indices, overall reading strategies and reading purposes, which were reported introspectively and validated retrospectively, for the text-reading context, while for the test context, they were only reconstructed retrospectively in the students' discussion of their test answers.

The main originality of the current investigation is that it sought a quantification of the effect of the test context, primarily, the test item type, MC and OE, on the test performance, by employing an independent, test-free assessment of reading, namely, the reading of a text, which was given a score, as will be explained later, and then compared to the subjects' performance on the test. Such a contrast between reading across contexts, different item-types, and of different native language communities, reading native and foreign language texts, in two different reading contexts, was hoped to cast light on the quantity and the quality of the effect that the testing context, namely, the test and the item type, exercise on the reading performance and process of different readers, to examine the potential reader and native-language-based differences in reading native and foreign language texts and tests, and to glean insight into the implications of such findings to the construct validity of reading comprehension tests.

The Current Research

The currently reported paper, being only a part of an ongoing doctoral research, does not intend to cover all the foregoing issues, but will rather be confined to a discussion of the effect of the reading context, testing versus non-testing, and MC vs. OE, only on the reading performance, rather than process, of native and foreign language texts. Specifically, it will attempt to answer the following research questions:

1. Does the reading context, i.e. testing vs. non-testing exert a significant effect on the reading performance of subjects

a) when reading in the native language: Arabic or Hebrew.

b) when reading in English as a foreign language?

2. Does the test item type, namely, MC or OE have a significant effect on the performance of subjects, in the test contexts, in the native and foreign language reading?

3. Is there a difference in the extent of such an effect of context and item type between the two native language communities, reading in the native and in the foreign language?

Subjects

Subjects of the current study were university students enrolled in courses in English as a foreign language, at the pre-advanced and Advanced-1 levels of the department of foreign languages at Haifa university. There were a total of 62 subjects, 19 males and 43 females; 24 Arabic and 38 Hebrew native speakers. The age of subjects ranged from 18 to 38 years old, and they came from a variety of departments, to the exclusion of the department of the English language and literature. Because all university teaching is conducted in Hebrew, even

at the Arabic-language and literature department, information on each subject's major university field was not expected to be significant.

methodology and instrumentation

In order to investigate reading in the foreign language, in a testing and a non-testing context, two elicitation tasks were needed, namely, an English language text, and an English language test, i.e., a text plus a set of questions. In addition, since the researchers were interested in the reading of English as a foreign language, the texts were to be designed for readers of English as a foreign language. Such texts, controlled for readability, length, and neutrality of topic, were adapted from the 4-point English Bagrut Examination of the years 1983 and 1990, one of which took the form of a test. Two versions of questions were developed, Multiple-Choice(MC) and Open-Ended (OE). The OE questions were simply the stem version of the MC without the alternatives, however, being aware of the fact that items cannot perform equally well as MC and OE items, such a potential pitfall was heeded from the outset, and the items adopted were those to have proved successful in both versions. The items were pretested with a group of forty high-school students at the Yanni High-school of Kfar-Yassif, and items were analyzed for their facility and discrimination values, and were moderated accordingly.

Similar to the aforementioned procedure followed in the writing of the English elicitation tasks, the Hebrew and Arabic texts were supposed to meet the criterion of target Hebrew and Arabic native communities, respectively. This was achieved by adopting texts from "High Q" (1992), and Kiddum (1993), which are practice books in preparation for the psychometric test, including authentic, previously administered tests of native language

Hebrew and Arabic reading comprehension. The texts chosen were again controlled for their readability level, their length, and the neutrality of the topic discussed, so that reliance on background knowledge, in the comprehension of the text, is minimized. Two sets of MC and OE items were developed on one of the two texts, following the same procedure just described in the writing of the native language items.

Data Collection Procedure

As a teacher at the department of foreign languages at Haifa university, the researcher experimented with her own students. To be sure the students took the task seriously and found time for them, they were told, with the approval of the department chair person, that the tasks from one of the course requirements. Thus, students were invited to the language laboratory in pairs, and were given the following tasks to perform:

1. The Native Language Text Reading: Upon arrival at the language laboratory, subjects were seated individually, each with earphones, and were given the NL text to read. The text, like all other texts, was divided into five paragraphs, each marked, at its end, with an asterisk. Informants were instructed to read the text out-loud, and to think aloud the strategies they were using to make sense of the text, whenever they wished to. They were reminded, however, that the asterisk was to draw their attention to such an introspection task at the end of the paragraph, if they had not yet performed it. Students were asked to paraphrase the text as they were reading, to explain in their native language the kind of meaning the were getting out of the text. This mentalistic method of data collection, i.e., the introspection, think aloud procedure , was adopted despite the validity and reliability problems it suffers from, which are, as Faerch et al. (1980) claims are "the hazards of science" (p. 395), because such a method is capable of revealing aspects of language

behavior that are otherwise inaccessible, and because, as Ericsson et al. (1980) points out, can be both a reliable and valuable source of information provided that it is "elicited with care and interpreted with full understanding of the circumstances under which they were obtained" (p. 274).

2. The English Text Reading: Next, subjects were given a text in English to read. Here dictionaries were allowed, and students received the same instructions they did prior to reading the L1 text.

3. The Native Language Test Taking: Next, students were administered a test, and were assigned 45 minutes to complete it. Pair were given tests of the same method, either MC, or OE.

4. The English Test Taking: Subjects were given an English test to complete within an hour, with the help of the dictionary if necessary. Pairs were given tests of the same method.

5. The Structured Interview: After they had performed these tasks, the researcher carried out a structured interview with each of the subjects. Following Grotjahn (1987) recommendation for the use of "methods of controlled understanding of others" (p. 67), in order to reduce the subjective discretion in interpreting introspection data, the purpose of this interview was to provide an 'interpretive validity' of the informants' reports, by eliciting more information on the subjects reading behavior, in the different contexts. Contradictory reports were discarded.

6. The Pair Discussion: Following the interview, each pair of subjects, who had taken same native language tests using the same method, were given their tests back, and were asked to sit together and record a discussion of the answers they had provided to the native and foreign language tests. Students were informed to explain why they had given those specific answers and to try and convince one another of the merits of their own answers. This discussion was meant to provide the researcher with information on the strategies subjects had employed in answering test questions, since if asked to do so during the taking of the test, it would have had been at the expense of sacrificing a real-life testing situation.

Data Analysis

In order to examine the reading performance and process across the two reading contexts, we assigned each subject two reading comprehension scores, namely, the score he received on the test, MC or OE, and the score given to the paraphrase/translation part of the think-aloud protocol. As we previously explained, instructions prior to the introspection task, required of students, to provide on-line a paraphrase/translation of the text, i.e., what they understood from it. This section was evaluated against a 10-main idea criterion list, selected, on consensus, by the researcher and three other colleagues, to constitute the main ideas of the text. Four English, Hebrew, and Arabic experts, including the researcher, were given the reading texts and were asked to locate the main ideas which would, in aggregate, give a complete account of the content each of the texts. The responses of judges were compared, and those ten ideas which all four judges included in their main-idea lists, were selected. These ten main ideas were then used as the criterion against which the paraphrase/translation section was evaluated yielding, what was called, the Introspection Score.

In terms of scoring, no specific observations were noted in the scoring of the MC test, however, in the OE test, it is worth pointing out that scoring has focused on the reading comprehension aspect, while tolerating spelling, grammatical, or writing mistakes, i.e., if the answer could indicate a comprehension of the idea under question, it was considered correct. No partial points were given in either introspection or test scores, in order to enhance scorer reliability, and for the sake of fairness, especially, to those taking the MC version, which, by nature, does not allow for partially correct responses.

The reading *process* across the reading contexts was examined by probing the sorts of overall reading strategies and reading purposes that characterized the reading comprehension process in each context. In addition, the introspection protocols were used to follow the specific reading behavior readers were performing, and to check whether there exists any significant correlation between the behavior of different readers of different characteristics with the kind of meaning they constructed. Different reading behavior indices were recorded by counting the frequency of their appearance in the protocol, such as, regressions, two to three words backwards, rereading, going backwards in the text to read previous sentences, mispronunciation, which results in a confusion with another acceptable word, pausing, for more than two seconds within minimal syntactic units, pausing within single word boundaries, searching for word meaning, either in the dictionary, or a verbalized attempt at retrieving meaning from long-term memory, and exchanging words for others.

The *within-test* reading process was investigated, as previously mentioned, initially, in terms of overall reading strategies and purposes, as reported in the testee-researcher

interview, and then in the analysis of the sources of test errors, as reported in the subjects' pair discussion.

Results and Discussion

With regards the first research question, namely, whether the reading context, test vs. non-test, exerts a significant effect on the reading performance of subjects, the present study yielded such a significant difference between the reading performance of subjects across the two reading contexts. Although a significant correlation was found between the subjects' performance in the two reading contexts, namely, test and introspection, and in the reading of both native and foreign language texts and tests (correlation coefficient .5742, p. = .000 and .8801, p. = .000, respectively), the t-test yielded significant differences between the two scores in both languages, with the introspection scores being significantly and consistently higher than the test scores. The following table summarizes the means of performance within each context, and the difference in performance attributable to context.

Table 1

	Test score			Introspection score			test score-introsp. score			
	N	X	SD	N	X	SD	N	X	T-val	sig
Eng	62	6.13	2.20	55	7.42	1.88	54	-1.0]	-8.15	0.000
NL	62	8.10	1.72	55	9.16	1.03	54	-0.8]	-5.41	0.000

]* p= .019

This finding, namely, the significant difference between the performance of subjects, in the two reading contexts, and in both native and Foreign language reading, brings under much questioning the construct validity of the reading comprehension tests used, and the amount

of error variance the test items contribute to the performance of readers on reading comprehension tests, in both languages. In addition, the significantly higher reading performance subjects proved to demonstrate, in the non-test situation, in both languages, casts much doubt on the ethical facet of reading assessment, and the extent of fairness we, as testers, and test users, exercise towards the testees we are entrusted with evaluating.

Now regarding the relative effect of the test-reading context on the reading performance in the native and foreign language, as the above table shows, the difference in the reading performance that can be attributed to the reading context, was found significantly greater when reading the English foreign language than when reading in the native language (p.= .019). This finding points at the relative role that test items play in the test performance, which seems to be greater in the foreign language test performance than in the native language reading, apparently due to differences in language proficiency. Reading in the native language, which involves competent readers does not seem to be adversly affected by the amount of additional reading data provided by the questions. This is different from reading in the foreign language, where the reading proficiency of readers is not yet fully developed, and where the construction of the meaning of the input text relies, to a much greater extent, on the readers' comprehension of the questions as well as of the text. This differential role of the items between the two languages suggests that the weaker the reading proficiency of readers is, the more the effect that questions exercise on reading , and in turn, the weaker the construct validity of the reading test is. This conclusion, reinforces the suggestion made by Anderson et al. (1991), namely, the combination, for the construct validation of reading comprehension tests, of multiple sources of information, particularly, in this respect, information on the readers' level of proficiency in the language, which is likely to affect the kinds of strategies they employ in the performance on the reading test, and

consequently on the construct validity it proves to have. In order to back this conclusion up, further investigation into the effect of the test reading context, across different ability levels in the native and foreign language, is in order.

Now regarding the effect of such a context difference within each of the two native language communities investigated, we found that this effect was significant with each of the two native language communities, reading both native and foreign language texts and tests. The following table summarizes the mean of performance and the difference in performance attributable to context, within each group.

Table 2

	NL Test score			NL Introspection score			NL test score-introsp. score			
NL	N	X	SD	N	X	SD	N	X	T-val	sig
Heb	38	7.6]*	1.82	31	9.16	1.13	31	-1.2]*	-5.66	0.000
Ar	24	8.8]	1.19	24	9.17	0.92	23	-0.3]	-1.91	0.069
	Eng Test score			Eng Introspection score			Eng test score-introsp. score			
Heb	38	5.97	2.43	31	7.35	2.14	31	0.94	-4.90	0.000
Ar	24	6.38	1.79	24	7.50	1.53	23	-1.26	-7.47	0.000

]* p<.005

The above table also shows that in terms of native language differences in the reading performance of subjects across contexts, the difference in the mean of performance between the Arabic and Hebrew native speakers performing on their native language tests, was found significant (p= .001), with the Arabic native speakers outperforming the Hebrew native speakers. In addition, the effect of the reading context, namely, the difference in performance between reading a test and reading a text, was found significantly greater for the Hebrew native speakers, than for the Arabic native speakers (p= .000). Regarding reading in the

foreign language, no significant differences were found between the two native language communities.

These findings strengthen the previously made statement concerning the amount of error that the reading test context introduces into the reading performance of both native language communities, reading native language and foreign language texts and tests. In addition, Having controlled for the readability of the two native language test-texts, the significantly superior reading performance of Arabic native speakers on the native language test, over their Hebrew language counterparts, and at the same time, the significantly greater effect of the test context, for Hebrew native speakers, suggests that the items introduce more error into the reading of Hebrew native speakers, than they do for the Arabic native speakers, and that this is what accounts for the superior performance of the Arabic native readers over the Hebrew natives. However, because reading a test is not molded, merely, by the test items, and because reading in the native language, in particular, and reading, in general, is influenced by a host of linguistic, contextual and attitudinal variables, the explanation of this difference, in the performance of Arabic native speakers on the Arabic language test, was also sought in the process aspect of reading, in the adaptation of the reading process to the reading purpose, in the nature of the Arabic language orthography, and in the attitudes Arabic native speakers hold of reading in Arabic, which stems from the Diglossic situation of the Arabic language, and which they seem to transfer to reading in the foreign language. A detailed discussion of these issues is beyond the scope of this paper.

As to the last question, which inquires the effect of the item type, or test method, on the performance of subjects in the test context, no significant difference in the performance of subjects, was found attributable to the test method. However, it was found that the relative

effect of the reading context on the performance of subjects on the foreign language tests, across the two testing methods, was significantly greater when the test used the multiple-choice method than when the method was open-ended. This led to the conclusion that multiple-choice as a testing method had a stronger effect on the reading performance than open-ended questions.

Moreover, there was a tendency for performance on the multiple-choice item type to give higher reading performance in the native language test, while lower reading scores in the foreign language test. The following table summarizes the findings concerning the test method variable.

Table 3

	NL Test score			NL Introspection score			NL test score-introsp. score			
	N	X	SD	N	X	SD	N	X	T-val	sig
MC	37	8.14	1.72	33	9.27	0.94	32	-0.88	-5.46	0.000
OE	25	8.04	1.74	22	9.00	1.15	22	-0.86	-2.66	0.014
	Eng Test score			Eng Introspection score			Eng test score-introsp. score			
MC	37	5.97	2.27	32	7.56	1.70	31	-1.3]*	-7.79	0.000
OE	25	6.36	2.12	23	7.17	2.12	23	-0.7]	-3.87	0.000

]* p= .03

The results reported above, reinforce the major claim made by the current study regarding the effect that the test weilds on test performance, and reinstantiates the implications of that to the construct validity of reading comprehension tests. Moreover, the differential relative effect of the reading contexts across the two testing methods, especially in foreign language reading, strengthens the above issue even further and suggests the inclusion, in the construct validation of reading comprehension tests, of information on item type characteristics.

To sum up, the results of the current study reinforce the amount of error that the test, regardless of which item type, introduces into the reading test performance, for both native language communities, reading in both the native and foreign language reading contexts, and reinstantiates, the implications of that to the construct validity of reading comprehension tests. The test reading context resulted in a significantly different performance than that achieved in the paraphrase/translation context in both languages, and the different item types proved to play a differential role in native and foreign language reading across the two reading contexts. This reinstates the error that the reading test contributes to the test performance, in native and foreign language reading contexts, and brings under much doubt the construct validity of reading comrehension tests.

REFERENCES

Alderson, J.C. 1990. Testing reading comprehension skills, Part one, Journal of Reading in a Foreign Language, vol. 6, pp. 425-38.

Alderson, J.C. and Lukmani, Y. 1989. Cognition and Reading: cognitive levels as embodied in test questions, Journal of Reading in a Foreign Language, vol. 5, pp. 253-70.

Anderson, N., Bachman, L., Perkins, K. and Cohen, A. 1991. An exploratory study into the construct validity of a reading comprehension test: triangulation of data sources, Language Testing, vol. 8, pp. 41-66.

Bachman, L.F., Davidson, F., and Milanovic, M. 1996. The use of test method characteristics in the content analysis and design of EFL proficiency tests, Language Testing, vol. 13, pp. 125-150.

Cohen, A.A. 1984. On taking language tests: what the students report, Language Testing, vol. 1, pp. 70-81.

Ericsson, K.A. and Simon, H.A. 1980. Verbal reports as data, Psychological Review, vol. 87, pp. 215-151.

Faerch, C. and Kasper, G. 1987. Introspection in Second Language Research, Multilingual Matters Ltd.

Farr, R. , Pritchard, R. and Smitten, B. 1990. A description of what happens when an examinee takes a multiple-choice reading comprehension test, Journal of Educational Measurement, vol. 37, pp. 209-26.

Freedle, R. and Kostin, I. 1993. The prediction of TOEFL reading item difficulty: implications for construct validity, Language Testing, vol. 10, pp. 131-171.

Grotjahn, R. 1987. On the methodological basis of introspective methods, in Faerch, C and Kasper, G. (eds.), Introspection in Second Language Research, Multilingual Matters Ltd.

Messick, S.A. 1989. Validity, in Linn, R.L. (ed.), Educational Measurement, (3rd edition), New York, American Council on Education. Macmillan Publishing Company.

Nevo, N. 1989. Test-taking strategies on multiple-choice test of reading comprehension, Language Testing, vol. 6, pp. 201-15.

Spolsky, B. 1990. Social aspects of individual assessment, in de Jong, J.H.A.L. and Stevenson, D.K. (eds.), Individualizing the Assessment of Language Abilities, Multilingual Matters Ltd., pp. 3-15.

Storey, P. 1994. Reading and problem solving: the interaction of individual differences and task types, in Proceedings of the International Language Education Conference,

Department of the curriculum studies, University of Hong Kong, pp. 121-147.

בחינת בגרות באנגלית, 4 יחידות לימודת קיץ תשמ"ג 1983.

בחינת בגרות באנגלית, 4 יחידות לימוד, קיץ תש"ן 1990.

High Q הכנה לבחינה הפסיכומטרית, מהדורה ראשונה, אוקטובר 1992.

كيدوم , ڤحضير لامتحانات البسيخومتري , طبعه جديدة ١٩٩٣ , דפוס "הד" בע"מ.

BEST COPY AVAILABLE